University of Zurich[UZH]

Zurich Open Repository and Archive
University of Zurich
University Library
Strickhofstrasse 39
CH-8057 Zurich
www.zora.uzh.ch

Year: 2010

Inkrementelle Koreferenzanalyse für das Deutsche

Klenner, M ; Tuggener, D ; Fahrni, A

Abstract: Es wird ein inkrementeller Ansatz zur Koreferenzanalyse deutscher Texte vorgestellt. Wir zeigen anhand einer breiten empirischen Untersuchung, dass ein inkrementelles Verfahren einem nichtinkrementellen überlegen ist und dass jeweils die Verwendung von mehreren Klassifizierern bessere Resultate ergibt als die Verwendung von nur einem. Zudem definieren wir ein einfaches Salienzmass, dass annähernd so gute Ergebnisse ergibt wie ein ausgefeiltes, auf maschinellem Lernen basiertes Verfahren. Die Vorverarbeitung erfolgt ausschliesslich durch reale Komponenten, es wird nicht - wie so oft - auf perfekte Daten (z.B. Baumbank statt Parser) zurückgegriffen. Entsprechend tief sind die empirischen Ergebnisse. Der Ansatz operiert mit harten linguistischen Filtern, wodurch die Menge der Antezedenskandidaten klein gehalten wird. Die Evaluierung erfolgt anhand der Koreferenzannotationen der TüBa-D/Z.

Posted at the Zurich Open Repository and Archive, University of Zurich
ZORA URL: https://doi.org/10.5167/uzh-39611
Conference or Workshop Item
Accepted Version

Originally published at:
Klenner, M; Tuggener, D; Fahrni, A (2010). Inkrementelle Koreferenzanalyse für das Deutsche. In: KONVENS 2010, Saarbrücken, 6 September 2010 - 8 September 2010, 37-46.

Inkrementelle Koreferenzanalyse für das Deutsche

Manfred Klenner
Institut für Computerlinguistik
Universität Zürich
Schweiz
klenner@cl.uzh.ch

Don Tuggener
Institut für Computerlinguistik
Universität Zürich
Schweiz
tuggener@cl.uzh.ch

Angela Fahrni
HITS gGmbH
Heidelberg
Deutschland
Angela.Fahrni@h-its.org

Abstract

Es wird ein inkrementeller Ansatz zur Koreferenzanalyse deutscher Texte vorgestellt. Wir zeigen anhand einer breiten empirischen Untersuchung, dass ein inkrementelles Verfahren einem nicht-inkrementellen überlegen ist und dass jeweils die Verwendung von mehreren Klassifizierern bessere Resultate ergibt als die Verwendung von nur einem. Zudem definieren wir ein einfaches Salienzmass[1], dass annähernd so gute Ergebnisse ergibt wie ein ausgefeiltes, auf maschinellem Lernen basiertes Verfahren. Die Vorverarbeitung erfolgt ausschliesslich durch reale Komponenten, es wird nicht - wie so oft - auf perfekte Daten (z.B. Baumbank statt Parser) zurückgegriffen. Entsprechend tief sind die empirischen Ergebnisse. Der Ansatz operiert mit harten linguistischen Filtern, wodurch die Menge der Antezedenskandidaten klein gehalten wird. Die Evaluierung erfolgt anhand der Koreferenzannotationen der TüBa-D/Z.

1 Einleitung

Empirische Untersuchungen zur Koreferenzresolution gehen oft von Vereinfachungen aus. Die gängigsten Idealisierungen sind:

- Ausklammerung der Anaphorizitätsentscheidung (*true mentions only*)
- Nutzung einer Baumbank
  - perfekte Bäume
  - perfekte funktionale Information (grammatische Funktionen)
  - perfekte morphologische Kategorisierungen

[1] Die Schweizer Rechtschreibung kennt kein scharfes s.

Oft wird von sog. *true mentions* ausgegangen, man verwendet also nur diejenigen Nominalphrasen (NPen), die gemäss Gold Standard in einer Koreferenzmenge sind. Die meisten nicht-pronominalen Nominalphrasen sind aber nicht-anaphorisch. Das Problem, diese von den anaphorischen zu unterscheiden, wird ausgeklammert. Empirische Werte unter dieser Setzung liegen etwa 15% zu hoch, vgl. (Klenner & Ailloud, 2009).

Es wird heutzutage nahezu ausschliesslich mit statistischen Methoden (inkl. *machine learning*) gearbeitet. Anaphernresolution wird dann meist als paarweise Klassifikationsaufgabe aufgefasst. Es werden unter anderem folgende Merkmale (*features*) beim Trainieren eines supervisierten Klassifizierers verwendet: die grammatischen Funktionen der NPen, ihre Einbettungstiefe im Syntaxbaum, die Wortklasse und die morphologischen Merkmale der Köpfe der NPen. Falls, wie im Falle der TüBa-D/Z, eine Baumbank mit diesen Spezifikationen vorhanden ist, ist die Versuchung gross, auf reale Komponenten zu verzichten und diese Goldstandardinformation zu nutzen.

Man findet auch wirklich kaum ein System, das vollständig auf realen Modalitäten beruht, d.h. das diese Merkmale mit einem Parser und einer Morphologie zu bestimmen versucht. Die Frage, wo wir heute bei der Lösung der Probleme im Bereich der Koreferenzresolution stehen, ist daher nicht einfach zu beantworten. Es ist deswegen auch schwierig, verschiedene Systeme zu vergleichen: Wir (die Autoren dieses Beitrags eingeschlossen) idealisieren gewissermassen aneinander vorbei. Dabei sind mittlerweile auch für das Deutsche Chunker und Parser verfügbar (z.B. (Sennrich et al., 2009)).

Der vorliegende Beitrag ist vor allem dazu gedacht, eine Baseline für das Deutsche

aufzustellen, die ausschliesslich auf realen Vorverarbeitungskomponenten beruht: Gertwol, GermaNet, Wortschatz Leipzig, Wikipedia und einem Dependenzparser, der auf einer manuell konstruierten Grammatik basiert, jedoch eine statistische Disambiguierung durchführt. Wir haben verschiedene empirische Szenarien durchgespielt und sind zu dem Ergebnis gekommen, dass ein inkrementelles, mit harten Filtern operierendes System zur Koreferenzanalyse (Nominal- und Pronominalanaphern) die besten Ergebnisse liefert. Zur paarweisen Klassifikation der Antezedens-Anapher-Kandidaten verwenden wir Timbl, ein ähnlichkeitsbasiertes Lernverfahren. Wir zeigen, dass ein einzelner Klassifizierer schlechtere Ergebnisse liefert als der Einsatz von vielen, auf die einzelnen Anapherntypen (Personal-, Possessiv-, Relativ- und Reflexivpronomen, sowie Nomen) zugeschnittenen Klassifizierern. Dies gilt für beide Varianten: die nicht-inkrementelle Variante und die inkrementelle (die mit einer Client-Server-Architektur realisiert ist). Wir experimentieren mit verschiedenen Merkmalen, wie sie in der Literatur diskutiert werden. Darüberhinaus erlaubt unser inkrementelles Modell aber neue, koreferenzmengenbezogene Merkmale.

Eigentlich dazu gedacht unsere Baseline zu definieren, stellte sich das von uns empirisch definierte Salienzmass als beinahe ebenbürtige, vor allem aber wesentlich einfachere und schnellere Alternative zu dem Einsatz eines Klassifizierer heraus.

Eine Demoversion unseres inkrementellen Systems ist unter http://kitt.cl.uzh.ch/kitt/cores/ verfügbar.

2 Modelle zur Koreferenzresolution

Was ist das beste Modell zur Koreferenzanalyse? Inkrementell oder nicht-inkrementell; ein Klassifizierer oder viele Klassifizierer; salienzbasiert oder mittels maschinellem Lernen?

3 Salienz

Zur Modellierung von Salienz existiert eine Reihe unterschiedlicher, aber miteinander verwandter Ansätze. (Lappin & Leass, 1994) legen manuell Gewichte für grammatikalische Funktionen fest (das Subjekt erhält den höchsten Wert) und begünstigen neben kurzer Entfernung zwischen Antezedenskandidat und Anapher die Parallelität von grammatikalischen Funktionen. Unser Mass (siehe weiter unten) kann als eine empirische Variante dieser Idee interpretiert werden. (Mitkov, 1998) adaptiert und erweitert das Verfahren. Er modelliert z.B. mit dem Givenness-Merkmal das Thema-Rhema-Schema (z.B. ob eine NP die erste im Satz und somit Thema ist). Weiter wird u.a. bestimmt, ob eine NP im Titel oder in einer Überschrift vorkommt (Section Heading Preference), wie oft sie im Text vorhanden ist (Lexical Reiteration) und ob sie von einer Menge bestimmter Verben fokussiert wird (Indicating Verbs). Auch die Definitheit der NP wird berücksichtigt.

Seit diesen Pionierarbeiten hat sich in Bezug auf die Salienzmodellierung nicht mehr viel getan. Vielmehr wurde mit der Auswertung und Aufsummierung der Salienzwerte experimentiert (z.B. statistische Verfahren etwa (Ge et al., 1998)). Später wurden die Merkmale in Ansätze mit maschinellem Lernen übernommen.

Wir haben mit einem rein korpusbasiertem Salienzmass experimentiert, wobei die Salienz einer NP durch die Salienz der grammatischen Funktion gegeben ist, die die NP innehat. Die Salienz einer grammatischen Funktion GF ist definiert als die Anzahl der Anaphern, die als GF geparst wurden, dividiert durch die Gesamtzahl der Anaphern. Es handelt sich also um folgende bedingte Wahrscheinlichkeit:

P(Y realisiert GF|Y ist eine Anapher).

Insgesamt 13 grammatische Funktionen (oder Dependenzlabel) erhalten so eine Salienz, darunter: Adverbiale (adv), Appositionen (appo), direkte und indirekte Objekte, das Subjekt. Wie erwartet ist das Subjekt mit 0.106 am salientesten, gefolgt vom direkten Objekt.

Unser Salienzmass kann in Verbindung mit unserem inkrementellen Modell folgenderweise verwendet werden: Für jedes Pronomen wird das salienteste Antezedens ermittelt. Dieses ist über die Salienz der grammatischen Funktion des Antezedenskandidaten und - bei gleicher Funktion - über seine Nähe zum Pronomen (i.e. der Anapher) eindeutig und sehr schnell bestimmbar. Für Nominalanaphern eignet sich dieses Mass nur bedingt, da Nomen im Gegensatz zu den meisten Pronomen ja nicht in jedem Fall anaphorisch sind (Salienz kennt aber keinen Schwellenwert, was zur Auflösung viel zu vieler Nomen führen würde).

3.1 Nicht-inkrementelles Modell

Die meisten Ansätze zur Koreferenzresolution sind sequentiell: erst wird ein Klassifizierer trainiert, dann werden alle (Test-)Paare gebildet und dem

Klassifizierer in einem Schritt zur Klassenbestimmung (anaphorisch, nicht-anaphorisch) vorgelegt. In einem solchen nicht-inkrementellen, paarweisen Verfahren sind alle Entscheidungen lokal – jede Entscheidung ist völlig unabhängig von bereits getroffenen Entscheidungen. Dies führt u.a. dazu, dass nachträglich ein Koreferenzclustering durchgeführt werden muss, bei dem alle Paare in Koreferenzmengen zusammengeführt werden und nach Möglichkeit Inkonsistenzen beseitigt werden. Denn es entstehen (nahezu unvermeidbar) z.B. nicht kompatible Zuordnungen von transitiv verknüpften Ausdrücken. So ist in der Sequenz 'Hillary Clinton .. sie .. Angela Merkel' jeder Name mit dem 'sie' kompatibel, die beiden Namen selbst hingegen nicht (da zwei nicht-matchende Eigenname i.d.R. nicht koreferent sein können). Trifft ein lokal operierender Klassifizierer für jedes Pronomen-Namen-Paar jedoch eine positive Entscheidung, entsteht via Transitiviät eine Inkonsistenz (H. Clinton und A. Merkel sind dann transitiv koreferent). Diese muss das nachgeschaltete Clustering beseitigen, indem eine der beiden Koreferenzentscheidungen (der beiden Pronomen-Namen-Paare) rückgängig gemacht wird. Es gibt verschiedene Clusteringansätze: *best-first* (das wahrscheinlichste Antezedens), *closest-first* (das am nächsten liegende), *aggressive merging* (alle positiven werden verknüpft). Man kann diesen Clusteringschritt aber auch als Optimierungsproblem auffassen, vgl. (Klenner, 2007) und (Klenner & Ailloud, 2009), wobei der Klassifizierer Gewichte liefert und linguistische motivierte Constraints die Optimierungsschritte beschränken. Unser Clusteringverfahren arbeitet mit einem Algorithmus aus dem Zero-one Integer Linear Programming, dem Balas-Algorithmus (Balas, 1965). Dabei werden die Antezedens-Anapher-Kandidaten aufsteigend nach ihrem Gewicht geordnet (Minimierung der Kosten ist das Ziel) und von links nach rechts aufgelöst: solange keine Constraints verletzt werden, wird jedes Paar mit Kosten kleiner 0.5 koreferent gesetzt. Verglichen mit einem *aggressive merging* gewinnt man so 2-4 % F-Mass.

Ein weiteres Problem der paarweisen Klassifizierung (ob inkrementell oder nicht-inkrementell) ist das Fehlen globaler Kontrolle. Obwohl z.B. Possessivpronomen in jeden Fall anaphorisch sind (Ausnahmen sind in der TüBa-D/Z sehr rar), kann man dies dem Klassifizierer nicht als Restriktion jeder gültigen Lösung vorschreiben. Es tritt sehr häufig der Fall ein, dass Promomen vom Klassifizierer kein Antezedens zugewiesen bekommen (d.h. kein Paar kommt über die Schranke von 0.5%). Dies kann man im nicht-inkrementellen Modell nachträglich durch eine Forcierung von Koreferenz reparieren, indem man den besten (d.h. am wenigsten) negativen Kandidaten als Antezedens nimmt. Im Falle eines inkrementellen Modells kann eine solche Bindungsforderung direkt eingelöst werden.

3.2 Inkrementelles Modell

Im Gegensatz zum nicht-inkrementellen Ansatz sind bei einem inkrementellen Ansatz die entstehenden Koreferenzmengen sofort verfügbar, Klassifikationsentscheidungen werden nicht auf einzelne Antezedenskandidaten, sondern auf die gesamte Koreferenzmenge, bzw. einen prototypischen Stellvertreter bezogen. So etwa im obigen Beispiel: entscheidet der Klassifizierer beispielsweise, dass 'sie' eine Anapher zu 'Hillary Clinton' ist, also die Koreferenzmenge [Hillary Clinton, sie] eröffnet wird, dann wird die NP 'Angela Merkel' nicht wie im nicht-inkrementellen Fall mit 'sie' verglichen, sondern mit 'Hillary Clinton' oder gar einem virtuellen Stellvertreterobjekt der Koreferenzmenge, das die Eigenschaften der ganzen Koreferenzmenge repräsentiert. So liefert 'Hillary Clinton' eine semantische Restriktion (Person, weiblich), aber keine morphologische. Obgleich 'sie' morphologisch ambig ist (z.B. Singular und Plural), kann es im Zusammenspiel mit der Information 'weiblich' auf den Singularfall in der dritten Person restringiert werden.

Ein weiteres Problem nicht-inkrementeller Ansätze ist, dass zu viele negative Beispiele generiert werden (vgl. (Wunsch et al., 2009) wo dieses Problem mittels Sampling gelöst werden soll). Dies führt zu einer Verzerrung des Klassifizierers, er erwirbt eine Präferenz zur negativen Klassifikation. Auch dies kann mit einem inkrementellen Modell abgemildert werden, denn pro Koreferenzmenge muss nur einmal verglichen werden; die restlichen Elemente der Koreferenzmenge sind nicht erreichbar. Dies reduziert die Menge der Paare insgesamt - sowohl der positiven als auch der negativen (siehe die Verhältnisangaben im Abschnitt 'Experimente').

Die Paargenerierung wird durch Filter restringiert. Neben den naheliegenden morphologischen Bedingungen (z.B. Person-, Numerus- und Genuskongruenz bei Personalpronomen), gibt es semantische Filter basierend auf GermaNet und

1	for	i=1	to laenge(I)
2		for	j=1 to laenge(C)
3			r_j := repräsentatives, legitimes Element der Koreferenzmenge C_j
4			Cand := Cand $\oplus$ r_j if kompatibel(r_j, m_i)
5		for	k= laenge(P) to 1
6			p_k:= k-tes, legitimes Element des Puffers
7			Cand := Cand $\oplus$ p_k if kompatibel(p_k, m_i)
8	if	Cand	= {} then P := P $\oplus$ m_i
9	if	Cand	$\neq$ {} then
10		ordne	Cand nach Salienz oder Gewicht
11		b	:= dasjenige e aus Cand mit dem höchsten Gewicht
12		C	:= erweitere(C,b,m_i)

Figure 1: Inkrementelle Koreferenzresolution: Basisalgorithmus

Wortschatz Leipzig bei den Nominalanaphern. Die semantischen Filter sind sehr restriktiv, so dass viele 'false negatives' entstehen, was eine recht tiefe Obergrenze (*upper bound*) generiert (ein F-Mass von 75.31%, eine Präzision von 81.58% und eine Ausbeute von 69.95%). Zwei Nomen sind semantisch kompatibel, wenn sie synonym sind, oder eines das Hyperonym des anderen ist. Nicht-kompatible Paare werden ausgesondert (das Prinzip von harten Filtern). Unsere Experimente haben gezeigt, dass restriktive Filter besser sind als laxe oder gar keine Filter.

Abbildung (Figure) 1 gibt den Basisalgorithmus zur inkrementellen Koreferenzresolution wieder. Seien I die chronologisch geordnete Liste von Nominalphrasen, C die Menge der Koreferenzmengen und P ein Puffer, in dem NPs gesammelt werden, die nicht anaphorisch sind (aber vielleicht als Antezedens in Frage kommen); m_i sei die aktuelle NP und $\oplus$ repräsentiere Listenverkettung (genauer 'Hinzufügen eines Elements'). Für jede NP werden anhand der existierenden Koreferenzpartition und dem Puffer (einer Art Warteliste) Kandidaten (Cand) generiert (Schritte 2-7), denen dann entweder von einem Klassifizierer oder über die Salienz ihrer grammatischen Funktion ein Gewicht zugewiesen wird (Schritt 10). Der Antezedenskandidat b mit dem höchsten Gewicht wird ausgewählt und die Koreferenzpartition wird um m_i erweitert (Schritt 11 und 12). Je nachdem, was b ist, heisst das, dass die Koreferenzmenge von b um m_i erweitert wird, oder dass eine neue Koreferenzmenge bestehend aus m_i und b eröffnet wird. Falls keine Kandidaten gefunden wurden, wird m_i gepuffert, da es zwar selbst nicht anaphorisch ist, aber als Antezedens für nachfolgende NPen verfügbar sein muss (Schritt 8). Pronomen und (normale) Nomen müssen in einem Fenster von 3 Sätzen gebunden werden (dies ist die Bedeutung von 'legitim' in den Zeilen 3 und 6), Eigennamen können auch weiter zurück (auf Eigennamen) referieren (auch durch diesen Filter werden 'false negatives' produziert und auch hier gilt, dass ein liberales Setting zu schlechteren empirischen Werten führt).

Die Kompatibilität zweier NPen (Schritte 4 bzw. 7) ist POS-spezifisch. Zwei Personalpronomen müssen z.B. im Numerus, Genus und Person kongruieren, während zwei Nomen nur im Numerus übereinstimmen müssen ('der Weg' .. 'die Strecke'), jedoch semantisch kompatibel sein müssen. Im Moment beschränkt sich dies auf eine GermaNet-Abfrage (Synomyme und Hyponyme sind erlaubt) und Wortschatz Leipzig (Synonyme).

Mit Blick auf die Paargenerierung (beim Machine Learning) lässt sich sagen: Die Anzahl der generierten Paare verringert sich bei unserem Verfahren um eine durch die Anzahl und Grösse der Koreferenzmengen bestimmten Betrag. Je mehr NPen bei geringer Anzahl von Koreferenzmengen (aber grösser Null) in diesen Koreferenzmengen gebunden sind, desto weniger Paare werden generiert (siehe Abschnitt 'Experimente' für konkrete Zahlen). Der Grund: ein Anapherkandidat m_i wird nur mit einem Element jeder Koreferenzmenge gepaart. Sei bei einem Fenster von 3 Sätzen die Kardinalität von I = 10 (also 10 NPen) und C bestehe aus einer einzigen Koreferenzmenge, die 6 Elemente aus I enthalte, dann wird m_{10} (die linear gesehen letzte zu integrierende NP) nur mit 5 NPen statt mit 9 gepaart: einem Element der Koreferenzmenge und den 3 gepufferten. Auf diese Weise reduziert sich auch die Anzahl der negativen

Beispiele, da für jede Koreferenzmenge (egal, ob m_i dazu gehört oder nicht) ja immer nur ein Glied betrachtet wird.

Das inkrementelle Verfahren gibt uns neue, koreferenzmengebezogene Merkmale an die Hand. Wir können daher weitere, bislang in der Literatur nicht verwendete Merkmale definieren:

- stammt der Antezedenskandidat aus dem Puffer oder einer Koreferenzmenge?
- Anzahl der gefundenen Kandidaten
- Anzahl der momentanen Koreferenzmengen
- Neueröffnung einer Koreferenzmenge oder Erweiterung einer bestehenden?

  (die folgenden Features beziehen sich auf die ausgewählte Koreferenzmenge)

- wieviele Nomen hat die Koreferenzmenge?
- Kardinaliät der Koreferenzmenge

Unsere empirischen Ergebnisse zeigen, dass der Nutzen dieser Merkmale vom Anapherntyp abhängt.

4 Vorverarbeitung

Die Vorverarbeitung dient der Extraktion linguistischer (morphologischer, syntaktischer und semantischer) Beschreibungen, die beim Filtern von Paarkandidaten bzw. als Merkmale beim Machine Learning verwendet werden. Wir verwenden Gertwol (Lingsoft, 1994), den TreeTagger (Schmid, 1994), GermaNet (Hamp & Feldweg, 1997), Wortschatz Leipzig (http://www.wortschatz.uni-leipzig.de), Wikipedia und den Parser Pro3GresDe (Sennrich et al., 2009).

Neben der Bestimmung des Lemmas und der morphologischen Kategorien, führt das Morphologieanalysetool Gertwol auch eine Nomendekomposition durch, was sehr hilfreich ist, da Komposita oft nicht in GermaNet gefunden werden, jedoch nach der Dekomposition ihre semantische Klasse anhand des Kompositakopfes oft richtig bestimmt werden kann.

Numerus, Genus und Person sind wesentlich für das Ausfiltern von sicheren negativen Paaren. Es gibt jedoch das Problem der Unterspezifikation und Ambiguität, z.B. bei den Pronomen 'sie', 'sich' und 'ihr'.

Die Named-Entity Erkennung ist musterbasiert und benutzt eine grosse Liste von Vornamen (53'000), wobei das Geschlecht zum Vornamen bekannt ist. Wir haben zudem aus der deutschen Wikipedia alle Artikel, deren Suchterm ein Mehrwortterm ist, extrahiert (z.B. 'Berliner Sparkasse') und, falls verfügbar, die zugehörige Wikipediakategorie (und diese, falls möglich, auf GermaNet abgebildet). GermaNet bzw.Wortschatz Leipzig liefern Synonyme und Hyponyme.

Pro3GresDe, ein hybrider Dependenzparser für das Deutsche, der eine handgeschriebene Grammatik mit einer statistischen Komponente zur Disambiguierung kombiniert, liefert u.a. die grammatische Funktion von NPen.

5 Experimente

Die folgenden empirischen Ergebnisse beruhen auf der TüBa-D/Z, einer Baumbank, die ebenfalls mit Koreferenzannotationen versehen wurde (24'000 annotierte Sätze in unserer Version), vgl. (Naumann, 2006). Als Klassifizierer verwenden wir das ähnlichkeitsbasierte Lernverfahren TiMBL (Daelemans et al., 2004)). Als Evaluationsmass wird nicht der MUC-Scorer verwendet, sondern der ECM bzw. CEAF aus (Luo, 2005). Beim CEAF erfolgt zuerst eine Alinierung von Koreferenzmengen des Gold Standard mit den vom System gefundenen. Präzision ist dann die Anzahl der richtigen Elemente pro alinierter gefundener Menge durch die Anzahl der gefundenen, bei der Ausbeute wird entsprechend durch die Anzahl der tatsächlichen Elemente der Gold Standard Menge geteilt. Der CEAF ist ein strenges Mass, da u.U. auch in nicht alinierbaren Mengen richtige Paare existieren. Der Vorteil: der CEAF ermittelt die Güte der Koreferenzmengenpartitionierung.

Unser Ansatz ist filterbasiert, d.h. Paare, die die Filter nicht passieren, werden als negativ klassifiziert. Darunter sind viele *false negatives* und zwar vor allem im Bereich der Nominalanaphern. Dies sei am Beispiel der ersten 5'000 Sätze illustriert. In Abbildung (Figure) 2 werden die oberen Schranken (*upper bound*) mit und ohne Filter aufgelistet. Die tatsächliche Performanz (im gewählten Fold) des inkrementellen Systems mit mehreren Klassifizierern ist: F-Mass = 53.86%, Präzision = 54.64% Ausbeute = 53.09% Hätten wir z.B. eine perfekte Nominalanphernresolution (die erste Zeile), dann könnten das System unter sonst unveränderten Bedingungen (die anderen Anapherntypen werden weiterhin vom System aufgelöst), maximal 62.61% F-Mass erreichen (mit Filter); ohne Filter wären

	Ohne Filterung			Mit Filterung		
	F-Mass	Präzision	Ausbeute	F-Mass	Präzision	Ausbeute
Nomen	72.70	69.53	76.17	62.61	63.70	61.55
Personalpronomen	60.42	62.05	58.88	58.86	60.64	57.19
Relativpronomen	56.25	57.91	54.68	55.97	57.65	54.39
Possessivpronomen	56.06	57.35	54.82	55.81	57.18	54.51
Reflexivpronomen	55.68	57.11	54.32	54.16	55.64	52.77
Gesamt	-	-	-	75.31	81.58	69.95
System	-	-	-	53.86	54.64	53.09

Figure 2: CEAF-Werte bei perfekten Einzelklassifikatoren des inkrementellen Systems für die ersten 5000 Sätze (in %). Wie gut wäre das System, wenn es einzelne Wortklassen perfekt auflösen wurde? Mit Filterung heisst, dass nur die Paare perfekt ausgelöst werden, die entsprechende Filter passieren. Ohne Filterung bedeutet, dass alle gemäss Gold Standard positiven Paare der jeweiligen Wortklasse perfekt aufgelöst werden (die anderen Wortklassen werden vom realen System verarbeitet). Der Unterschied zwischen mit und ohne Filterung bezeichnet die Güte der Filter pro Wortklasse, die Unterschiede zum System, wie sehr die imperfekte Auflösung der Wortklasse das System drückt.

Verfahren	F-Mass	Präzision	Ausbeute
Nicht-inkrementell, ein Klassifizierer	44.04	55.60	36.48
Nicht-inkrementell, mehrere Klassifizierer	49.35	53.67	45.69
Inkrementell, ein Klassifizierer	50.66	52.54	48.93
Inkrementell, mehrere Klassifizierer	52.79	52.88	52.70
Salienz	51.41	52.03	50.82

Figure 3: CEAF-Werte der fünfachen Kreuzvalidierung (in %)

es 72.70%. Dies zeigt zweierlei. Nominalanaphern sind tatsächlich das Problem, wie die 9% Differenz zwischen Systemwert (53.86%) und perfektem Wert (62.61%) zeigt. Daneben erklärt die abermalige 10% Differenz zur perfekten Auflösung ohne Filter die insgesamt schlechte Performanz von 53.86%: insgesamt 19% Performanzverlust durch die Nominalanaphern. Im Vergleich zu den Reflexivpronomen. Hier ist die Differenz nur 0.3% (54.16% - 53.86%) zur perfekten Auflösung mit Filter und nur 1.8% zur perfekten Auflösung ohne Filter (55.68% - 53.86%).

Wir beschreiben nun unsere Experimente (fünffach kreuzvalidiert) zur Bestimmung des besten Ansatzes zur Koreferenzresolution, vgl. Abbildung (Figure) 3. Als Baseline dient ein nicht-inkrementelles Verfahren, das, wie alle Varianten, bzgl. Merkmalauswahl (features des Klassifizierers) optimiert wurde. Wir unterscheiden zwischen der Verwendung von einem und mehreren POS-spezifischen Klassifizierern.

Für das nicht-inkrementelle Verfahren mit nur einem Klassifizierer hat sich folgende Merkmalsmenge als am performantesten erwiesen: Salienz der NPen, grammatische Funktionen der NPen und ob diese parallel sind, Wortklasse der Köpfe der NPen und eine separate Kodierung der POS-Kombination, semantische Kompatibilität der NPen, Die Häufigkeit der NPen im Segment, ob der Antezedenskandidat der nächste kompatible zur Anapher ist, ob sich die NPen in direkter Rede befinden[2].

Dieser Ansatz (mit einem F-Mass von 44.04%) dient als erste Baseline. Die Verwendung mehrerer Klassifizierer (zweite Baseline) bringt eine Verbesserung um über 5% F-Mass (auf 49.35%). Das liegt daran, dass die Merkmale unterschiedlichen Einfluss auf die POS-spezifischen Klassifizierer haben. Distanz kann z.B. für Pronominalanaphern eingesetzt werden, bei den Nominalanaphern bringt sie nichts. Folgende Merk-

[2]Dass hier Distanz nicht verwendet wird, liegt daran, dass die Distanzmerkmale sich bei der Auflösung nominaler Anaphern als schlecht erwiesen haben. Da relativ viele nominale Paare bewertet werden müssen, schadet Distanz einem Einzelklassifizierer, der alle Arten von Anaphern bewerten muss.

malsmengen haben sich als am effektivsten herausgestellt:

- Nominalanaphern: Häufigkeiten der NPen im Segment, grammatische Funktionen der NPen, semantische Kompatibilität, Definitheit des Antezedenskandidaten, ob sich der Anaphernkandidat in direkter Rede befindet und ein Indikator für den Filter, der das Paar generiert hat (Stringmatch, GermaNet oder Wortschatz Leipzig).
- Personalpronomen: Distanz in Sätzen und Markables, Salienz der NPen, Einbettungstiefe der NPen, Wortklasse der Köpfe der NPen.
- Relativpronomen: Distanz in Markables, Salienz der NPen, grammatische Funktion der NPen, Einbettungstiefe des Anaphernkandidaten, ob der Antezedenskandidat der nächste kompatible zur Anapher ist.
- Reflexivpronomen: Distanz in Markables, Salienz des Antezedenskandidaten, Wortklasse der Köpfe der NPen und Kodierung der POS-Kombination, grammatische Funktionen, Einbettungstiefe der NPen, ob der Antezedenskandidat der nächste kompatible zur Anapher ist.
- Possessivpronomen: Salienz der NPen, Distanz in Sätzen, Einbettungstiefe der NPen, grammatische Funktionen der NPen und ob diese parallel sind, Wortklasse des Kopfs des Anaphernkandidaten, ob der Antezedenskandidat der nächste kompatible zur Anapher ist.

Wie oben erwähnt wurde, können im inkrementellen Modell Merkmale definiert werden, die sich auf die (entstehenden) Koreferenzmengen beziehen. Insgesamt hat sich die Verwendung dieser Merkmale als ambivalent herausgestellt: Nicht für alle Klassifizierer sind sie hilfreich. Bei der Verwendung nur eines Klassifizierers werden im inkrementellen Modell drei der erwähnten Merkmale verwendet: Kardinalität der Koreferenzmenge, Anzahl der Nomen in der Koreferenzmenge, ob eine neue Koreferenzmenge eröffnet wird. Ansonsten werden die gleichen Merkmale wie im nicht-inkrementellen Modell verwendet. Der Unterschied zu Baseline 2 ist mit 1.3% gering, doch spürbar.

Bei der Verwendung mehrerer Klassifizierer haben die Koreferenzmengen bezogenen Merkmale nur einen signifikanten Einfluss auf die Klassifizierer der Personal- und Possessivpronomen. Die Anzahl vorhandener Koreferenzmengen wird bei beiden verwendet, die Kardinalität der Koreferenzmenge zusätzlich bei den Personalpronomen. Ansonsten werden auch hier die gleichen Merkmale wie im nicht-inkrementellen Verfahren verwendet.

Bei der Verwendung mehrerer Klassifizierer werden, gegenüber Baseline 2, fast drei Prozentpunkte F-Mass dazugewonnen (52.79% vgl. mit 49.35%). Die Verbesserungen, die durch die Verwendung mehrerer Klassifizierer erreicht werden, entsteht v.a. durch einen Anstieg der Ausbeute. Auffallend ist, dass das Verhältnis von Präzision zu Ausbeute im inkrementellen Modell ausgeglichener ist als im nicht-inkrementellen.

Bezüglich Laufzeit ist festzuhalten, dass die nicht-inkrementellen Verfahren nicht nur mehr negative, sondern auch mehr positive Instanzen generieren, da alle Mentions aus den Koreferenzmengen für die Paargenerierung zugänglich sind. Diese zusätzlichen positiven und negativen Instanzen erhöhen die Laufzeit beträchtlich. Im letzten Fold (etwa 5000 Sätze) der Kreuzvalidierung z.B. generiert das nicht-inkrementelle Modell für das Training 23024 positive und 109344 negative Instanzen. Das inkrementelle Modell hingegen erstellt nur 10957 positive und 76955 negative Paare. Das entspricht bei den positiven Instanzen einer Reduktion von über der Hälfte, bei den negativen Instanzen um rund 30%. Da alle Verfahren die gleichen Filter verwenden, gehen in den inkrementellen Ansätzen keine *true mentions* verloren. Die Reduktion entsteht dadurch, dass Paare nur mit einem Element der jeweiligen Koreferenzmengen gebildet werden. Auch die Client-Server-Architektur des inkrementellen Modells beschleunigt die Laufzeit, da die TiMBL -Klassifizierer nicht für jede Klassifikation neu gestartet werden müssen.

Die letzte Zeile von Abbildung 3 gibt das Resultat der rein salienzbasierten Variante des inkrementellen Ansatzes wieder. Es schneidet erstaunlich gut ab und liegt mit 51.41% um 1.4% unter der Bestmarke. Diese gute Performanz bei der Einfachheit der Implementierung und der im Vergleich enorm kurzen Laufzeit, sind gute Argumente gegen die aufwändigere Implementation von Machine Learning Ansätzen. Dazu kommt, dass die Optimierung von Merkmalsmengen in Machine Learning Ansätzen einerseits nötig, andererseits aber auch zeitintensiv und die Auswirkungen einzelner Merkmalsetzungen unvorhersehbar ist. Die erzielten Verbesserungen aufgrund von Mu-

tationen der Merkmalsmengen können ausserdem linguistisch oft nur schwer begründet werden, resp. entziehen sich der Intuition. Ein Argument für die Verwendung von ML Verfahren ist aber die Behandlung von Bridging Anaphern, die in unserem salienzbasierten Verfahren nicht aufgelöst werden.

6 Literaturdiskussion

Die Arbeit von (Soon et al., 2001) ist ein prototypisches, oft reimplementiertes (Baseline-)Modell zur Anaphernresolution, das auf paarweiser Klassifikation und statistischen Verfahren basiert.

Eines der wenigen inkrementellen Modelle ist (Yang et al., 2004). Im Gegensatz zum vorliegenden Modell gibt es in diesem Ansatz für's Englische jedoch nur ein einziges echtes koreferenzmengenbezogenes Merkmal: die Anzahl der Elemente einer Koreferenzmenge.

Es gibt einige wenige Arbeiten zur Koreferenzresolution für das Deutsche, die meisten nutzen die Koreferenzannotation der Baumbank TüBa-D/Z. Uns ist kein System bekannt, das basierend auf realen Vorverarbeitungskomponenten sowohl Pronominal- als auch Nominalanaphernresolution modelliert. Die sehr aufschlussreiche Untersuchung von (Schiehlen, 2004) ist ebenfalls auf Pronominalanaphern beschränkt, zeigt aber wie tief die empirischen Werte tatsächlich liegen, wenn man reale Komponenten verwendet statt einer Baumbank.

(Versley, 2006) hat - auf der Basis einer Teilmenge der TüBa-D/Z - zahlreiche Experimente zur Nominalanaphernresolution durchgeführt (z.B. verschiedene statistische Masse um z.B. Selektionsrestriktionen zu modellieren). Sein Befund, dass wenn Information aus GermaNet verfügbar ist, diese dann statischer Information überlegen ist, hat uns dazu inspiriert, GermaNet durch Wikipedia und Wortschatz Leipzig zu komplementieren und auf statistische Berechnungen zu verzichten.

Neben GermaNet und einem pattern-basierten Ansatz, verwenden (Goecke et al., 2008) Latent Semantic Indexing bei der Nominalanaphernauflösung. Die empirische Analyse erfolgt anhand eines kleinen, von den Autoren eigens annotierten Korpus.

Modelle für Pronominalanaphern werden in einer Reihe von Arbeiten aus Tübingen diskutiert. Die empirischen Ergebnisse basieren auf Goldstandardinformation, so wird z.B. in (Wunsch et al., 2009) eine perfekte Morphologie und funktionale Information der TüBa-D/Z Baumbank verwendet. Diese Arbeit versucht das Problem der Übergenerierung von negativen Beispielen durch Sampling zu lösen. Das vorliegende inkrementelle Modell kann als Alternative dazu aufgefasst werden. Die Reduktion von Trainingsinstanzen ist ein natürlicher Nebeneffekt unseres inkrementellen Verfahrens.

7 Zusammenfassung und Ausblick

Es wurde ein Verfahren zur Koreferenzresolution für das Deutsche vorgestellt, das von realen Verarbeitungsmodalitäten ausgeht und sowohl Pronominal- als auch Nominalanaphern behandelt. Wir können festhalten, dass ein filterbasiertes inkrementelles Verfahren auf der Basis anaphernspezifischer Klassifizierer am besten arbeitet. Überraschenderweise ist der Abstand zu einem einfachen salienzbasierten System gering.

Die empirischen Werte sind mit knapp 52.79% F-Mass nicht berauschend. Schuld daran sind Fehler in den Annotationen der TüBa-D/Z (fehlende Annotationen bei Pronomen und matchenden Named Entities), Fehler beim Vorverarbeiten (z.B. Morphologie) und die Unterspezifikation im Bereich der Nominalanaphern (z.B. GermaNet-Lücken). Nominalanaphern bleiben die grosse Herausforderung.

Unser inkrementelles Verfahren ermöglicht eine natürliche Reduktion zu lernender Beispiele beim Vektorgenerieren, es ermöglicht uns darüberhinaus die Verwendung neuer Features wie z.B. die Anzahl der Koreferenzmengen. Nicht alle neuen Merkmalen helfen die Empirie zu verbessern, und unterschiedliche Anapherntypen profitieren von unterschiedlichen Merkmalen.

Unser Modell ist noch nicht ausgeschöpft. Verbesserungen erwarten wir u.a. im Bereich der Nominalanaphern. In jedem Fall aber liefert unser System eine nicht geschönte Baseline für die Koreferenzresolution des Deutschen.

Danksagung

Die Arbeiten zu unserem Projekt werden vom Schweizerischen Nationalfonds unterstützt (Nummer 105211-118108).

References

Egon Balas. 1965. An Additive Algorithm for Solving Linear Programs with Zero-one Variables. *Operations Research*, 13(4).

Walter Daelemans and J. Zavrel and K. van der Sloot and A. van den Bosch. 2004. *TiMBL:*

Tilburg Memory-Based Learner. Techn. Report. Tilburg University.

Niyu Ge and John Hale and Eugene Charniak. 1998. A Statistical Approach to Anaphora Resolution. *Proc. of the Sixth Workshop on Very Large Corpora*.

Daniela Goecke, Maik Stührenberg, Tonio Wandmacher. 2008. A Hybrid Approach to Resolve Nominal Anaphora. *LDV Forum*, 1(23).

Birgit Hamp and Helmut Feldweg. 1997. GermaNet—a Lexical-Semantic Net for German. *Proc. of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*.

Erhard W. Hinrichs and Katja Filippova and Holger Wunsch. 2005. A Data-driven Approach to Pronominal Anaphora Resolution in German. *Proc. of RANLP*

Manfred Klenner. 2007. Enforcing Consistency on Coreference Sets. *Proc. of the Ranlp*.

Manfred Klenner and Étienne Ailloud. 2009. Optimization in Coreference Resolution Is Not Needed: A Nearly-Optimal Zero-One ILP Algorithm with Intensional Constraints. *Proc. of the EACL*.

Shalom Lappin and Herbert J. Leass. 1994. An Algorithm for Pronominal Anaphora Resolution. *Computational Linguistics*.

Lingsoft. 1994. Gertwol. Questionnaire for Morpholympics. *LDV-Forum*, 11(1).

Xiaoqiang Luo. 2005. On Coreference Resolution Performance Metrics. *Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

Ruslan Mitkov. 1998. Robust Pronoun Resolution with Limited Knowledge. *Proc. of the ACL*. Montreal, Quebec, Canada.

Karin Naumann. 2006. *Manual for the Annotation of Indocument Referential Relations*. Tech. Report, Universität Tübingen.

Michael Schiehlen. 2004. Optimizing Algorithms for Pronoun Resolution. *Proc. of the 20th International Conference on Computational Linguistics*.

Helmut Schmid. 1994. Probabilistic Part–of–Speech Tagging Using Decision Trees. *Proc. of the Conference on New Methods in Language Processing*.

Rico Sennrich and Gerold Schneider and Martin Volk and Martin Warin. 2009. A New Hybrid Dependency Parser for German. *Proc. of the German Society for Computational Linguistics and Language Technology 2009 (GSCL 2009)*. Potsdam.

Wee Meng Soon and Hwee Tou Ng and Daniel Chung Young Lim. 2001. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*.

Holger Wunsch and Sandra Kübler and Rachael Cantrell. 2009. Instance Sampling Methods for Pronoun Resolution. *Proc. of RANLP*. Borovets.

Yannick Versley. 2006. A Constraint-based Approach to Noun Phrase Coreference Resolution in German Newspaper Text. *Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS)*.

Xiaofeng Yang and Jian Su and Guodong Zhou and Chew Lim Tan. 2004. An NP-Cluster Based Approach to Coreference Resolution. *Proc. of Coling*.